STAT 383C: Statistical Modeling I **Fall 2015**

Lecture 21 — November

Lecturer: Purnamrita Sarkar *Scribe: Carlos Zanini*

Note: These scribe notes have been slightly proofread and may have typos etc.

21.1 Bayesian Document Model

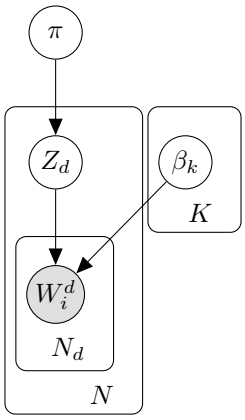
Consider the problem of classifying documents into topics based on their contents. The basic idea is: if a document contains, for instance, several words related to statistics (e.g., data, model, inference, likelihood), then it is likely that it belongs to Statistics topic.

We will see a model that allows more than one topic per document, but for now, the first model we are going to consider in this lecture supposes that a given document belongs to only one topic.

According to this model, a document is randomly generated as follows: first we randomly pick one topic Z_d, from the distribution over topics $\pi \sim \text{Dirichlet}(\lambda_1, ..., \lambda_K)$. The latent topic Z_d is just a distribution over words, and such distribution is fully specified by the probabilities $\beta = (\beta_1, ..., \beta_V)$ over the words in the vocabulary. Finally, the words for document d (denoted by $W_i^d,\ i = 1, ..., N_d$) are then randomly selected from the vocabulary with probabilities β. The words are the observed data, and all the other variables of the model are latent.

In summary, the model can be defined by the following set of statements:

- Z_d : topic of document $d \in \{1, ..., N\}$ (N: number of documents)
- $Z_d \in \{1, ..., K\}$ (K: number of topics)
- $\pi = (\pi_1, ..., \pi_K) \sim \text{Dir}(\alpha_1, ..., \alpha_K)$
  $\pi_k = P(Z_d = k), k \in \{1, ..., N\}$.
- W_i^d : word i in document $d \in \{1, ..., N\}$
  $W_i^d \in \{1, ...N_d\} \rightarrow$ vocabulary of the document d
- Distribution of W_i^d varies with the topic of document d.
  $P(W_i^d = w \mid Z_d = k) = \beta_{kw}, \quad \beta_k = (\beta_{k1}, ..., \beta_{zV})$
  V : total of words in the vocabulary
  $\beta_k \sim \text{Dir}(\lambda_1, ..., \lambda_{N_d}), \quad z \in \{1, ..., K\}$.



21.1.1 Full-conditional distributions for Gibbs sampler

Now we derive the full-conditional distributions for running a Gibbs sampler and get to the posterior distribution of the parameter of the model described before. Notice that all full-conditionals are analytically available and sampling from them is straightforward.

1.

$$\begin{aligned}\pi^{(t+1)} &\sim p(\pi \mid \{Z_d\}, \{W_i^d\}, \{\beta_k\}) \\ &= p(\pi \mid \{Z_d\}) \\ &= \text{Dir}(n_1 + \alpha_1, ..., n_K + \alpha_K),\end{aligned}$$

where $n_i = \sum_{d=1}^{N} 1(Z_d = i)$.

2.

$$\begin{aligned}\beta_k^{(t+1)} &\sim p(\beta_k \mid \{Z_d\}, \{W_i^d\}, \{\beta_{-k}\}, \pi) \\ &= p(\beta_k \mid \{Z_d\}, \{W_i^d\}) \\ &= p(\{W_i^d\} \mid \{Z_d\}, \beta_k) p(\beta_k \mid \{Z_d\}) \\ &= \left(\prod_{d=1}^{N} \prod_{i=1}^{N_d} p(W_i^d \mid Z_d, \beta_k) \right) \beta_{\lambda 1}^{\lambda_1} \dots \beta_{\lambda V}^{\lambda_V} \\ &= \left(\prod_{d=1}^{N} \prod_{i=1}^{N_d} \beta_{k W_i^d} \right) \beta_{\lambda 1}^{\lambda_1} \dots \beta_{\lambda V}^{\lambda_V} \\ &= \text{Dir}(m_{k1} + \alpha_1, ..., m_{kV} + \lambda_V),\end{aligned}$$

where $m_{kw} = \{(i, d) : W_i^d = w, z_d = k\}$, i.e. total number of occurrences of word w in documents of topic k.

3.

$$\begin{aligned} Z_d^{(t+1)} &\sim P(Z_d = k \mid \{Z_{-d}\}, \{W_i^d\}, \{\beta_k\}, \pi) \\ &= P(Z_d = k \mid W_1^d, ..., W_{N_d}^d, \beta_k, \pi) \\ &\propto p(W_1^d, ..., W_{N_d}^d \mid Z_d = k, \beta_k, \pi) P(Z_d = k \mid \beta_k, \pi) \\ &= \prod_{i=1}^{N_d} \beta_{kW_i^d} \times \pi_k \end{aligned}$$

Hence the full conditional distribution for Z_d is discrete given by

$$P(Z_d = k \mid \{Z_{-d}\}, \{W_i^d\}, \{\beta_k\}, \pi) = \frac{\prod_{i=1}^{N_d} \beta_{kW_i^d} \times \pi_k}{\sum_{k=1}^{N_d} \prod_{i=1}^{N_d} \beta_{kW_i^d} \times \pi_k}$$

21.1.2 Collapse Gibbs Sampler

Sometimes it is possible to marginalize the likelihood over a set of parameters, therefore reducing the number of nodes in the MCMC chain. This can save time during the simulation of the chains, since we have less nodes to sample from.

Here we exemplify how it works by showing the calculations for marginalizing out π from the full conditional of Z_d. Notice that this is enough to marginalize π out of the likelihood, since it does not appear in the full-conditional distribution of any of the β_k's.

$$\begin{aligned} P(Z_d = k \mid \{Z_d\}, \{W_i^d\}, \{\beta_k\}) &\propto \\ &\propto P(\{W_i^d\} \mid \{Z_{-d}\}, \{\beta_k\}) \times P(Z_d = k \mid \{Z_{-d}\}, \{\beta_k\}) \\ &= \prod_{d=1}^{N} p(W_i^d \mid Z_d, \beta_k) \times P(Z_d = k \mid \{Z_{-d}\}) \\ &\propto p(W_i^d \mid Z_d, \beta_k) \times \int \dots \int P(Z_d \mid \pi, \{Z_{-d}\}) p(\pi \mid \{Z_{-d}\}) d\pi_1 \dots d\pi_K \\ &= \prod_{i=1}^{N_d} \beta_{kW_i^d} \times \int \dots \int \pi_k \text{Dir}(\pi; n_1^{-d} + \alpha_1, ..., n_K^{-d} + \alpha_K) d\pi \\ &= \prod_{i=1}^{N_d} \beta_{kW_i^d} \int \pi_k \text{Beta}\left(\pi_k; n_k^{-d} + \alpha_k, \sum_{j \neq k} n_j^{-d} + \alpha_j\right) d\pi_k \\ &= \left(\prod_{i=1}^{N_d} \beta_{kW_i^d}\right) \frac{n_k^{-d} + \alpha_k}{\sum_j (n_j^{-d} + \alpha_j)} \end{aligned}$$

where $n_k^{-d} = |\{\ell \neq d : Z_\ell = k\}|, \quad k = 1, ..., K$.

It is possible to marginalize over β_k's too, but the calculations involved are not as simple as before and they will not be shown here. If you are interested, check out section 2.6 of [2].

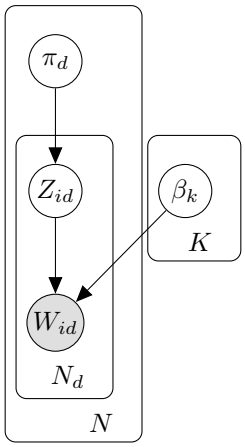
21.2 Latent Dirichlet Allocation

The latent Dirichlet allocation (LDA) model allows one document to belong to more than one topic. Actually, each word has a chance of being drown from a different topic and this is what makes a document to contain more than one topic. We now describe the model more precisely.

According to the LDA model, documents are randomly generated as follows. First we randomly select a distribution over topics. This distribution will tell us how often words will be sampled from topics $1, 2, ..., K$. Then for each word in the document, we randomly select a topic from it, using the distribution over topics that was already selected. Finally, we randomly generate a word according to the distribution over words for that specific topic.

You may consider taking a look at Figure 1 from [1], to help understanding the steps above.

The graphical representation of this model is very similar to the one presented for the previous model:



although presenting the following changes:

- Each word now has its own topic. And as a result a document no longer belongs to just one topic, but may possibly belong to many.

- π_d now represents the proportion of different topics for document d. Note there was only one π before.
  $\pi_d = (\pi_{d1}, ..., \pi_{dK}) \sim \text{Dir}(\alpha_1, ..., \alpha_K)$
  $\pi_{dk} = P(Z_{id} = k), k \in \{1, ..., K\}$.

- Each topic has a multinomial distribution over the vocabulary, e.g. topic k has vector β_k. $P(W_{id} = w \mid Z_{id} = k) = \beta_{kw}$, where $\beta_k \sim \text{Dir}(\lambda_1, ..., \lambda_V), \quad k \in \{1, ..., K\}$.

21.2.1 Full-conditional distributions for Gibbs sampler

Sampling π

$$\begin{aligned}\pi^{(t+1)} &\sim p(\pi_d \mid \{Z_i^d\}, \{W_i^d\}, \beta_k) \\ &= p(\pi \mid Z_1^d, ..., Z_{N_d}^d) \\ &= \text{Dir}(n_1^d + \lambda_1, ..., n_{N_d}^d + \lambda_{N_d})\end{aligned}$$

where n_i^d represents the numer of occurrencies of words drawn from topic i in document d.

Sampling β

$$\begin{aligned}\beta^{(t+1)} &\sim p(\beta_k \mid \{Z_i^d\}, \{W_i^d\}, \{\pi_d\}) \\ &= \text{Dir}(m_{1k} + \lambda_k, ..., m_{N^d k} + \lambda_{N^d}),\end{aligned}$$

where $m_{wk} = |\{(i,d) : W_{id} = w, Z_{id} = k\}|$. The expressions for the remaining full-conditionals are to be calculated in the next lecture.

References

[1] Blei, David. Probabilistic topic models. *Communications of the ACM, 55(4):7784, 201*, 2012.

[1] Resnik, P. and Hardisty, E. Gibbs Sampling for the Uninitiated. *CS-TR-4956 UMIACS-TR-2010-04 LAMP-TR-153*, June 2010.